Noise-Robust speech recognition based on acoustic features concatination.

Amrous Anissa Imen, Debyeche Mohamed

Faculty of Electronics and
Computer Sciences, USTHB
ALGERIA.
amrous_im@hotmail.fr, mdebyeche@gmail.com

ABSTRACT. In this paper we study the contribution of some auxiliary features to increase the robustness of an HMM-based speech recognition system in noisy environments. The front-end of the system combines features based on conventional Mel-Frequency Cepstral Coefficient (MFFC), and auxiliary information such as: pitch (fundamental frequency), energy and formants. Our HMM-based recognition system is implemented using the HTK toolkit and the ARADIGIT corpus. The obtained results show a significant improvement of the recognition system performance in noisy environment compared to standard system.

KEYWORDS: ASR system, HMM, MFCC, auxiliary Information.

.

1. Introduction

The standard Automatic Speech Recognition (ASR) systems are usually based on Hidden Markov Models (HMMs) and use generally cepstral-based features as acoustic parameters.The most powerful features currently used are the MFCCs (Mel Frequency Cepstral Coefficients), the LPC (Linear Prediction Coding) and the PLP (Perceptual Linear Predictive) [1]. However, these features are very sensitive to speech signal variability under real-life conditions [2, 3, 4]. The speech signal variability is mostly due to environmental factor (presence of noise) or to speaker characteristics (tiring, illness, gender …) and leads to different kinds of mismatch between acoustic features and acoustic models. This causes a reduction on the recognition rate under real-life conditions. The sensitivity of MFCC to noise motivates many authors to look for new parameters to make the acoustic models more robust. We can refer to Stephenson works [5] and Doss [6] who use within the framework of Dynamic Bayesian Networks *(*Dynamic Bayesian Networks are an alternative of HMMs*)* like auxiliary features: pitch, energy and Rate-Of-Speech (ROS). In addition, other works in the audio-visual domain have integrated the visual information in the acoustic recognition system [7] [8]. This work aims to integrate auxiliary knowledge sources into standard HMM-based ASR systems in order to make the acoustic models more robust to the speech signal variability [6]. The paper is set out as follows: section II describes the basic of standard hidden Markov model (HMM) based automatic speech recognition (ASR) systems. The section III presents baseline system description, whereas the proposed robust system is presented in section IV. In section V we show our experimental evaluations, and finally conclusions are presented in Section VI.

2. HMM based ASR system

The general architecture of standard HMM based ASR consists of three main components: parameters extraction, training, and recognition (Figure. 1).

2.1. Features extraction

Features extraction consists in converting the speech waveform signal into a parametric representation. This parametric representation is then used for training and recognition.

2.2. Training

Training an acoustic model on database training means estimate the parameters which characterize this acoustic model. In the case of Hidden Markov Model (HMM) these parameters are: the covariance matrix, the mean vector and the transition matrix. For that, the HMM models are initialized with Viterbi algorithm [9], then the Baum-Welch algorithm is called to train them [9].

2.3. Recognition

The recognition process calculates the likelihood between the observation sequences (the word to recognize) and all the acoustic models which are previously trained. The recognized word is the one which corresponds to the acoustic model leading to the maximum likelihood. This likelihood is computed using the Viterbi algorithm [9].

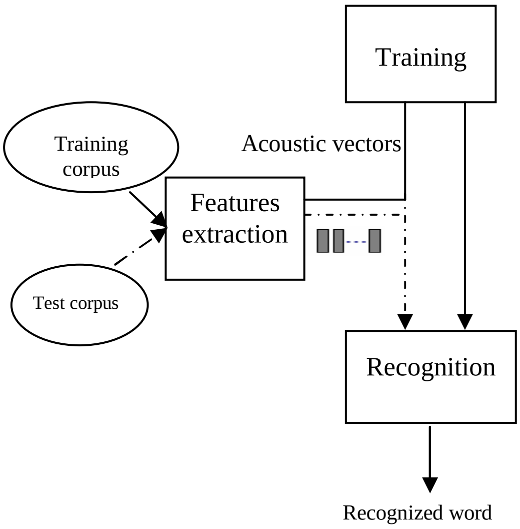

Figure1. *The general architecture of ASR system.*

3. Baseline ASR system

The baseline system is an isolated-word, speaker-independent system. This system uses cepstral features vectors as inputs. Thus, MFCCs extracted from the input speech signal, were generated the follow steps:

- Firstly, the speech signal is sampled at a frequency of16KHz
- In order to reduce the impact of the high frequencies, the speech signal is emphasized (the pre-emphasis coefficient is set to 0, 97 in our case).
- Since the speech signal is known as non-stationary, the signal analysis must be performed on a short-term basis. In this context, the speech signal is divided into a number of overlapping time windows of 25 ms with a frame period of 10 ms.
- For each analysis window, 12 Mel-Frequency Cepstral Coefficients (MFCCs) are calculated using a mel-scaled filterbank with 24 channels
- Then, the first (△) and second (△△) derivatives of MFCCs are appended to take into account the dynamic of the signal, making a total vector dimension of 36 (12 MFCC + 12 △ MFCC+ 12△△MFCC).

The HMM models are left-to-right HMMs with continuous observation densities. Each model consists of 3 states, in which, each *state* is modeled by 12 Gaussian mixture with a diagonal covariance matrice.

4. Proposed ASR sytem

Our proposed ASR system use as inputs a multivariate vectors composed of the MFCC vector (described in III) and new auxiliary features which are: pitch, energy and the first three formants. Those later were generated using Praat Toolkit [16] and then appended to the MFCC vectors by a simple concatenation. The theoretical background used to extract the new auxiliary features is as follow:

- Pitch : Its estimation is based on autocorrelation function [17].
- Formant frequencies : In this paper we choose to use the frequencies of the first three formants which are estimated from the maxima of the LPC spectrum model [18].
- Energy: The energy was computed by taking the logarithm of the windowed signal [13].

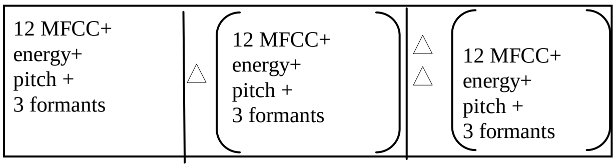

Figure 2. *Multivariate vectors composition.*

To complete the vectors, first (△) and second (△△) derivatives of multivariate vector are appended, making a total vector dimension of 51 (Figure. 2). Although that the concatenated features had different range of values, we did not use in the present implementation any special normalization

4. Experiments and results

4.1 Database

The speech database used in this work is the isolated ARADIGIT corpus [14]. It is composed of Arabic isolated digits from 0 until 9. This database is divided into the following corpuses:

- Train corpus: consisting of 1800 utterances pronounced by 60 speakers including the two genders, where, each speaker repeats the same digit 3 times.
- Test corpus: consisting of 1000 utterances pronounced by 50 speakers including the two genders, where, each speaker repeats the same digit 2 times.

This database was recorded in WAV format at 16 kHz of sampling frequency in clean conditions.

4.2 Experiments

Our experiments were developed using HTK package (Hidden Markov Toolkit) [13], from Cambridge University. With the aim to show the advantage of using auxiliary features in addition of cepstral features in speech recognition under real-life test conditions We carried out two sets of experiments, one for the baseline sytem and another for the proposed system.

The performance of these two systems in clean conditions and in adverse conditions (additive noise) has been studied. For the adverse conditions, we have corrupted the database with two kinds of noises, namely: factory noise and the pink noise. Both noises have been extracted from the NOISEX92 database [15] and added to the speech signal to achieve a Signal-to-Noise Ratio (SNR) of: 15 dB, 10dB and 5dB.

The acoustic models' training uses the clean speech database; the noises are only added for testing the recognition performance.

Word recognition rates obtained with both systems, baseline and proposed system in clean and noisy condition are summarized in Table I. The recognition results are given by the percent accuracy defined as:

$$accur = \frac{N - D - S - I}{N} \times 100 \qquad (1)$$

where N is the total number of units, D is the number of deletion errors, S is the number of substitution errors, I is the number of insertion errors.

Noise	SNR	Baseline system	Robust syetem
clean	35dB	99,45%	98,52%
Factory noise	15 dB	79,61%	84,78%
	10 dB	58,03%	80,30%
	5 dB	33,12%	60,33%
Pink noise	15 dB	77,95%	89,58%
	10 dB	57,10%	79,98%
	5 dB	30,63%	59,32%
Average	(35 dB to 5 dB)	62,27%	78,97%

Table1. *Comparative* speech *recognition results***.**

As it can be observed in Table I in clean conditions, recognition rates obtained with the proposed system are slightly worse than the baseline system (99.45%% vs. 98.52%). This can be explained by the fact that the additive features disturb the more reliable standard features.This disturbance did not only interfere at the recognition level, but also at the training level. Another reason which can explain this degradation of the recognition rate is the fact that the modeling of the system fusion vectors by a Gaussian mixture density may be an inappropriate choice to model the new vectors extended by the auxiliary additive features. Moreover the constraint of the diagonal covariance matrices is not suitable in presence of multivariate features as they are not uncorrelated. This motivates us to consider for further investigations new models with less constraint (e.g. neural network models).

It is worthy to note that in noisy conditions, the proposed system that includes additional features besides MFCCs clearly outperforms the baseline system. For example, with 5dB factory noise: 33.12%% vs. 60.33%, i.e., an improvement of 27.21% is noticed. It can be seen that, as the level of noise increases, the proposed system gains improvement in recognition accuracy over the baseline system. In case of pink noise, we noticed an increasing range of improvement from 11% to 29% according to SNR range 15 dB-5 dB respectively.

To summarize, by looking at the average accuracies in Table 1, one can observe that the performance of proposed system is better than that of the baseline system. This improvement is a consequence of the exploitation of the auxiliary features which allows the proposed system (proposed system) to have more information about the word to recognize under adverse conditions.

5. Conclusion

In this paper, we have studied the advantages of using auxiliary features to Arabic speech recognition system based on Hidden Markov Model. The auxiliary features are added to the state-of-the-art cepstral features, (MFCCs) by a simple concatenation of the two kinds of features. The obtained results suggest that auxiliary features could improve the ASR performance in noisy conditions. In fact, this inclusion yields an improvement of more than 29% of correct recognition rate in comparison with baseline system, under noisy conditions. Hence, we can conclude that the auxiliary features contain information which can be considered as complementary to the information provided by cepstral features (MFCC) and can be used to improve the speech recognition performance in noisy conditions.

6. References

[1] C. Lévy, G. Linarès and P. Nocera, “*Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems*”. Workshop on DSP in Mobile and Vehicular Systems, Nagoya - Japan, 2003.

[2] G.Baudoin, P.Jardin and al "*Comparison de techniques paramétrisation spectrale pour la reconnaissance vocale en milieu bruité*". Quatorzième colloque gretsi. September 1993.

[3] L. Mary and B. Yegnanarayana "*Extraction and representation of prosodic features for language and speaker recognition*". Speech Communication 50, 782–796.April 2008.

[4] H. Ezzaidi "Discrimination Parole/Musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques". Thèse de doctorat. Université du Québec à Chicoutimi. October 2002.

[5] T. A. Stephenson "Speech Recognition with Auxiliary". Thèse PHD. Ecole Polytechnique Fédérale de Lausane. May 2003.

[6] Mathew Magimai Doss "Using auxiliary sources of knowledge for automatic speech recognition". Thèse PHD; école Polytechnique Fédérale de Lausane. July 2005.

[7] P.Deleglise, A. Rogozan and M. Alissali "*Asynchronous integration of audio and visual sources in bi-model automatic speech recognition*". Proceedings of the VIII European Signal Processing Conference, Trieste (Italy). September 1996.

[8] Rodazana Alexandrina "Etude de la fusion des données hétérogènes pour la reconnaissance automatique de la parole audio-visuelle". Thèse doctorat de l'université d'Orsay Paris XI. July 1999.

[9] L.R. Rabiner "*A tutorial on hidden Markov models and selected applications in speech recognition*". Proc. of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[10] S.B. Davis and P. Mermelstein. "*Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences*". IEEE Trans. on Speech and Audio Processing, 28(4):357–366, 1980. 2.2

[11] L.R. Rabiner "*On the Use of Autocorrelation Analysis for Pitch Detection*". IEEE transaction on acoustics, speech, and signal processing,vol-25, 1. February 1977.

[12] S.B. Davis and P. Mermelstein. "*Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences*". IEEE Trans. on Speech and Audio Processing, 28(4):357–366, 1980. 2.2

[13] S. Young, J. Odell and al "The HTK Book Version 3.3". Speech group, Engineering Department , Cambridge University. April 2005.

[14] A. Amrouche "*Reconnaissance automatique de la parole par les modèles connexionnistes*" . Thèse de doctorat, faculté d'électronique et d'informatique, USTHB. 2007.

[15] A. P. Varga, H. J. M. Steeneken and al "*The NOISEX-92 study on the effect of additive noise on automatic speech recognition*". NOISEX92 CDROM, 1992.

[16] P. Boersma and Weenink, D. "*Praat: doing phonetics by computer*". From the web site: http://www.praat.org/ . 2008.

[17] L.R Rabiner " *On the Use of Autocorrelation Analysis for Pitch Detection*". IEEE Transaction on Acoustics, Speech, and Signal Processing 25, 1. 1977.

[18] Davis, S.B., Mermelstein, P.: "*Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences*". IEEE Trans. on Speech and Audio Processing 28(4), 357–366. 1980.